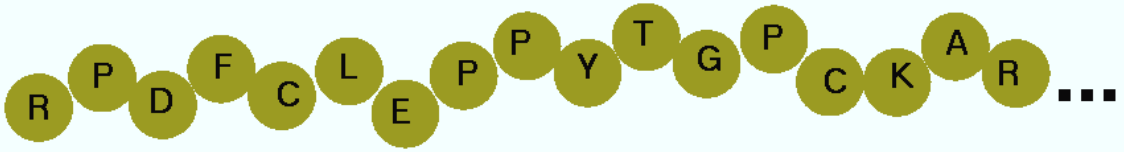
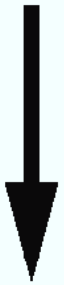
Calculating the Conformational Energy of Polypeptides and Proteins with ECEPPAK

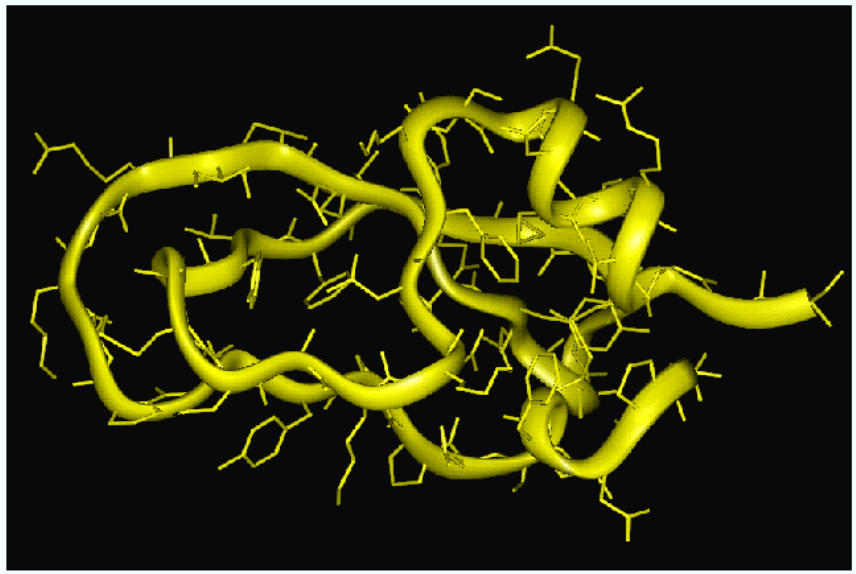
D. R. Ripoll

August 11, 2002

The Protein Folding Problem







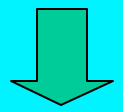
Protein Folding as a Problem of Global Optimization

- Anfinsen Experiments.
- From statistical mechanics:

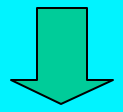
$$F = E - TS$$

Under equilibrium conditions, F is a minimum.

In its native state, a protein has a well-defined structure



Fluctuations around the native conformation are low



$$F \sim E$$

CTC

According to the ***thermodynamics hypothesis***, the protein folding problem is reduced to:

- **Finding a function, E_P**, to describe the potential energy (or force field).

- Search for the **global minimum** of the hypersurface defined by the potential energy, i.e., solve the ***multiple minima problem***

The Force Field

Hamiltonian: $H(\text{coordinates, moments}) = E_{total}$

The general expression of E_{total} (using flexible geometry) is:

$$E_{total} = K + V$$

where $$K = \sum_{\{atoms\}} 1/2\ [\mathbf{p}_i \cdot \mathbf{p}_i / m_i];$$

and

$$V = E_{bond} + E_{angle} + E_{dihedral} + E_{improper} + E_{Nonbonded}$$

$$E_{Nonbonded} = E_{electrostatic} + E_{van\ der\ Waals} + E_{loop} + E_{Hbond}$$

Force fields based on rigid geometry, like ECEPP assume fixed bond lengths and bond angles:

$$E_{bond} = E_{angle} = E_{improper} = 0$$

CTC

A modified Lennard-Jones 6-12 potential is used to compute the nonbonded interactions:

$$E_{\text{van der Waals}} = \Sigma_{ij}\, f\,(A^{kl} / r_{ij}^{12} - C^{kl} / r_{ij}^{6})$$

The electrostatic energy is computed as:

$$E_{\text{electrostatic}} = 332.0\ \Sigma_{ij}\, q_i\, q_j / D\, r_{ij}$$

The torsional energy (E_{dihedral}) is given by:

$$E_{\text{dihedral}} = \Sigma_{k}\ A(k)\,[1.+ ns(k)\cos(nb(k)\phi(k)]$$

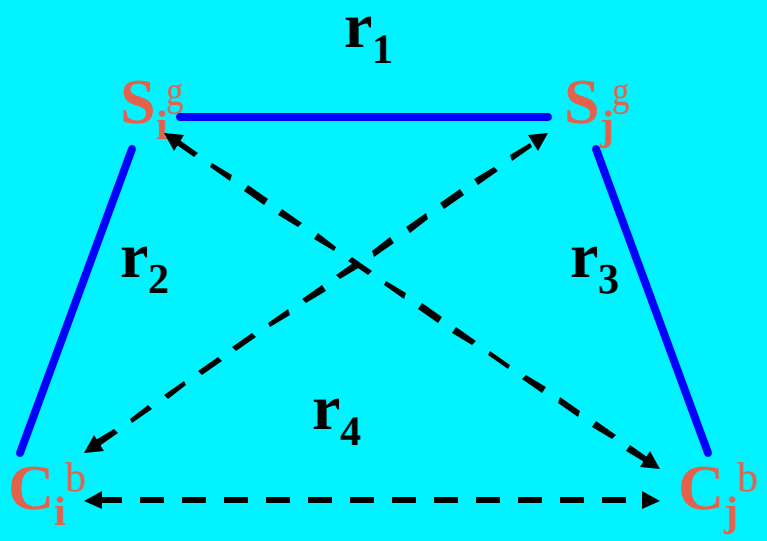
The hydrogen bond energy is:

$$E_{\text{Hbond}} = \Sigma\ A_{HX} / r_{ij}^{12} - B_{HX} / r_{ij}^{10}$$

CTC

Loop Closing Energy (E_{loop})

$$E_{loop} = A \sum^{3}_{i=1} (\mathbf{r_i} - \mathbf{r_{i,0}})^2 + B (\mathbf{r_i} - \mathbf{r_{i,0}})^2$$



Proline Residues:

Internal Conformational Energy depends on the pirrolidine ring geometry.

Continuum Solvation Models

• Provide precise solvation energies without introducing explicitly the solvent molecules.

• Computational costs are lower than those from simulations with explicit solvent molecules.

•Entropic solvent effects are included.

•Contributions to the solvation energy:

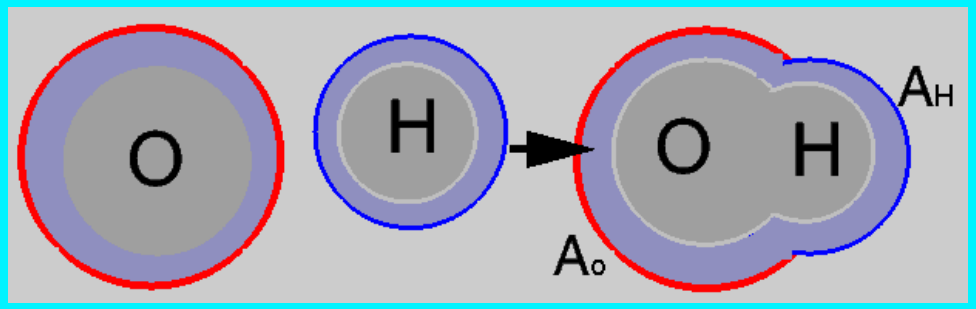
- costs of cavity creation

- dipolar alignment of solute and solvent

- dispersion energies

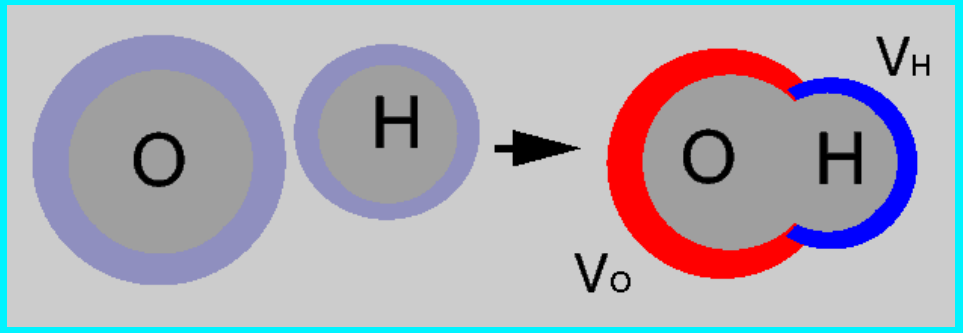
CTC

- Accessible surface area of solvent.



$$E_{ssurf} = A_H \beta_H + A_O \beta_O$$

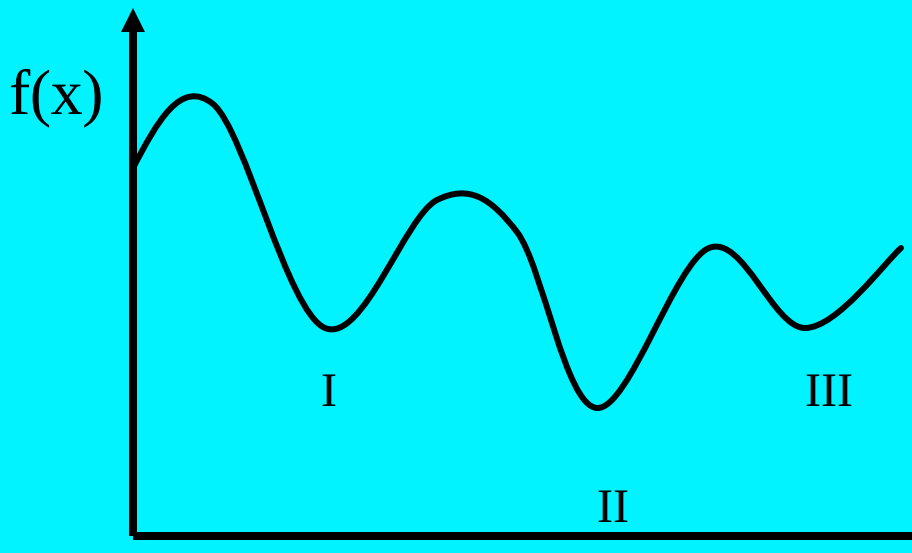
- Solvent-accessible volume



$$E_{svol} = V_H \upsilon_H + V_O \upsilon_O$$

where β_H , β_O, υ_H y υ_O are parameters that depend on the type of atoms.

Methods of local and Global Optimization



minimum of f(x)
$df/dx = 0$
$d^2f/dx^2 > 0$

I and III are local minima, II is the global.

- Local minimization finds the closest well, and depends on the starting point.
- Global minimization should locate II, independent of the starting point.

Global Minimization: intrinsic difficulties

- Realistic functions represent complicate surfaces in multi-dimensional spaces (***the multiple minima problem***).

- Lack of mathematical methods to solve the global optimization problem for general multi-variable functions.

Types of Global Optimization Methods

- Systematic search
  a- Grid search
  b- Build-up
  c- Self-Consistent Field Method.
  d- Diffusion Equation.
  e- Branch and Bound (Floudas).

CTC

Types of Global Optimization Methods (cont)

-Stochastic methods

a-Simple Random Search (Monte Carlo).
b- Important sampling MC.
c- Relaxation of the dimensionality.
d-Simulated annealing.
e- Monte Carlo with Minimization (MCM, EDMC).
f- Genetic algorithms.
g- Packing annealing.

The Electrostatically Driven Monte Carlo (EDMC) Method

- ***Basic assumptions:***

Folding of a polypeptide chain is driven primarily by two effects:

- Optimization of the electrostatic interactions.
- Thermal Effects.

- *Most relevant features of the EDMC method*

- Use of ***electrostatic predictions*** to generate new conformations.
- Use of randomly generated conformations.
- The ***backtrack mechanism*** is used to escape from specific regions of the conformational space when the search is trapped.

CTC

Features of the EDMC method

- A history of the search is kept by storing the set of accepted conformations and organizing them into clusters.

- Low-energy conformations from different clusters are used for random generation.

Generation of Conformations from electrostatic predictions

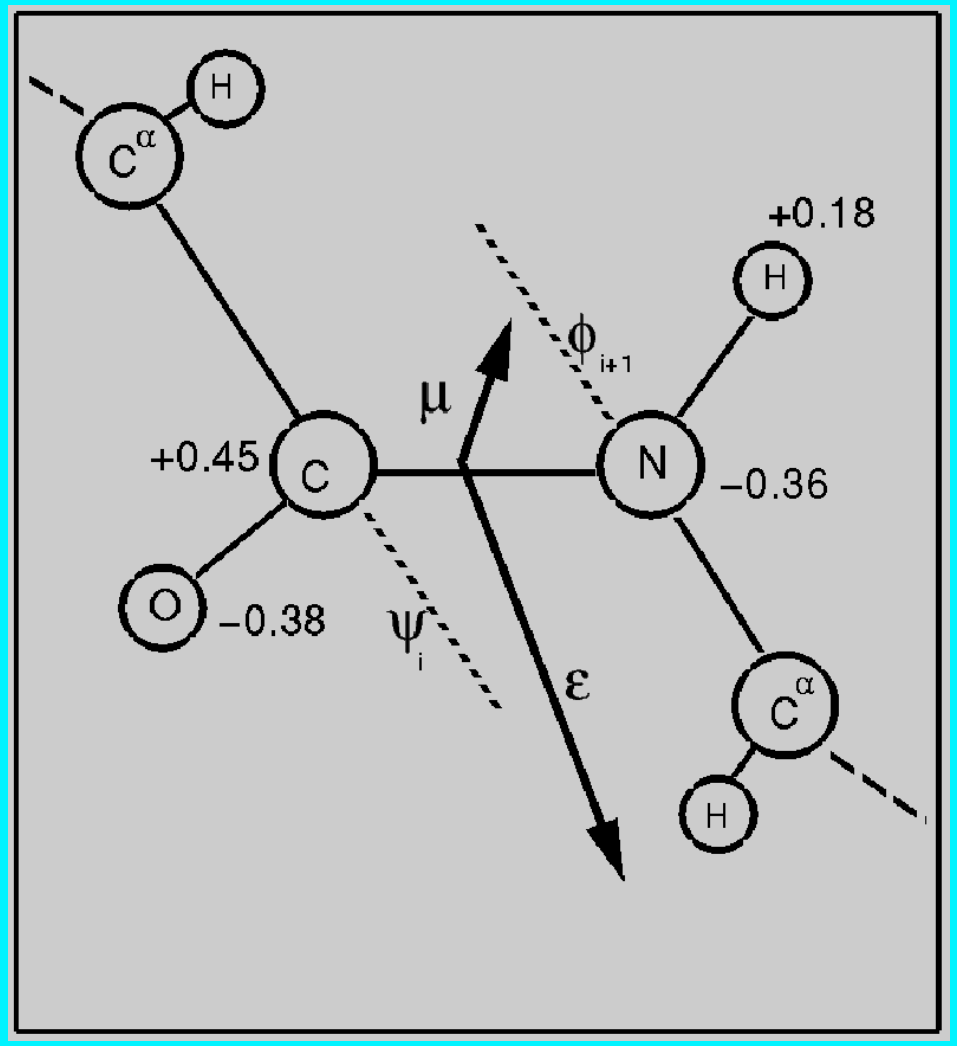
- The alignment of local dipoles with the electrostatic field generated by the rest of the molecule is analyzed for every newly accepted conformation.

- Local dipoles that are improperly oriented are used to determine possible movements by changes in the variables ϕ, ψ and and χ.

- From this analysis, a list of the possible changes with an estimate of the energy gain is generated.

CTC

Dipole Alignment

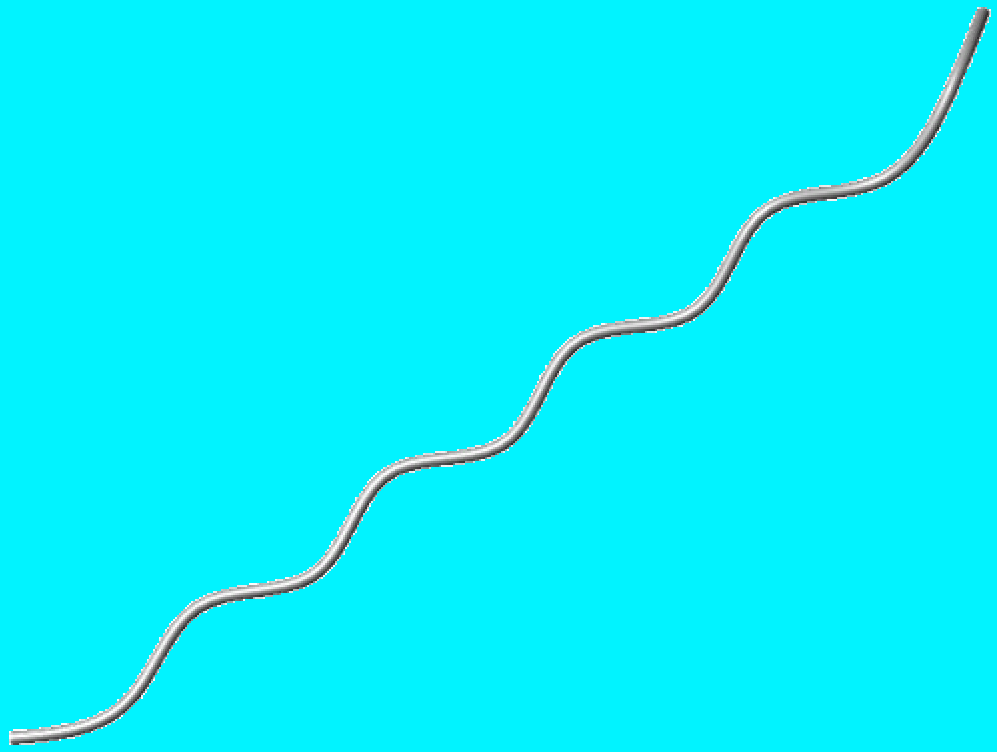


Features of the EDMC method

- Generation of Random Conformations

  Residues residues from a selected conformation are chosen randomly, then:

  1. A set of variables are randomly altered.
  2. Backbone variables are obtained using the ϕ-ψ map.
  3. Backbone variables are generated from a subset compatible with regular secondary structure.
  4. Use of pre-computed low-energy conformations of tripeptides.



Animation

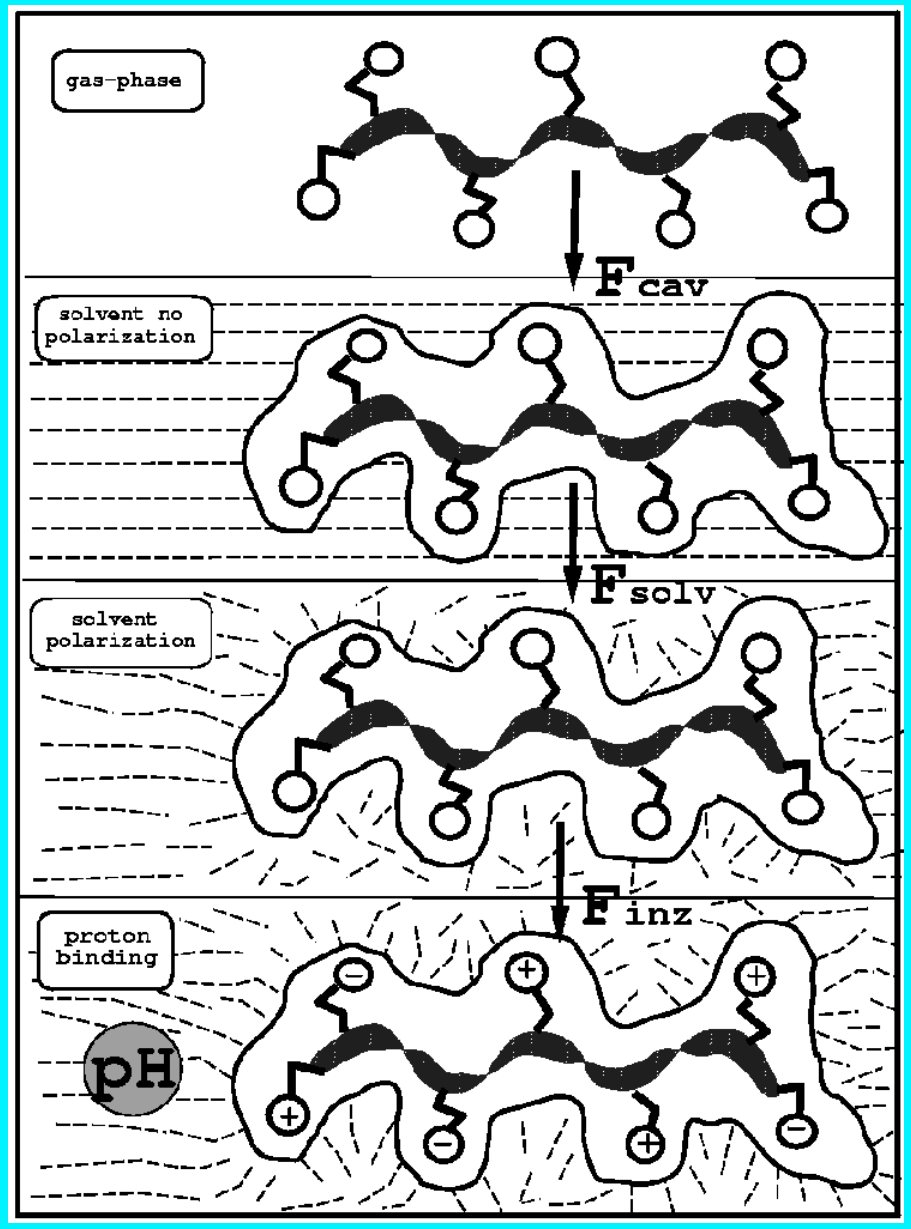
CTC

Conformation and Ionization Equilibrium

• The charge distribution on a molecule with ionizable groups depends on the environment conditions (e.g., pH and ionic strength) and on the conformation.

• The conformation, on the other hand, is defined itself by the electrostatic interactions.

•A reasonable calculation of the free energy of polypeptides must include the coupling between the states of ionization of charged groups and the conformation of a molecule, as a function of pH.


gas-phase
F cav
solvent no polarization
F solv
solvent polarization
F inz
proton binding
pH

•The free energy of polypeptides in aqueous solution is computed as:

$$E(\mathbf{r}_p, pH) = E_{int}(\mathbf{r}_p) + F_{vib}(\mathbf{r}_p) + F_{cav}(\mathbf{r}_p) + F_{solv}(\mathbf{r}_p) + F_{ionz}(\mathbf{r}_p, pH)$$

$E_{int}(\mathbf{r}_p)$ is the internal conformational energy of the molecule in the absence of solvent;

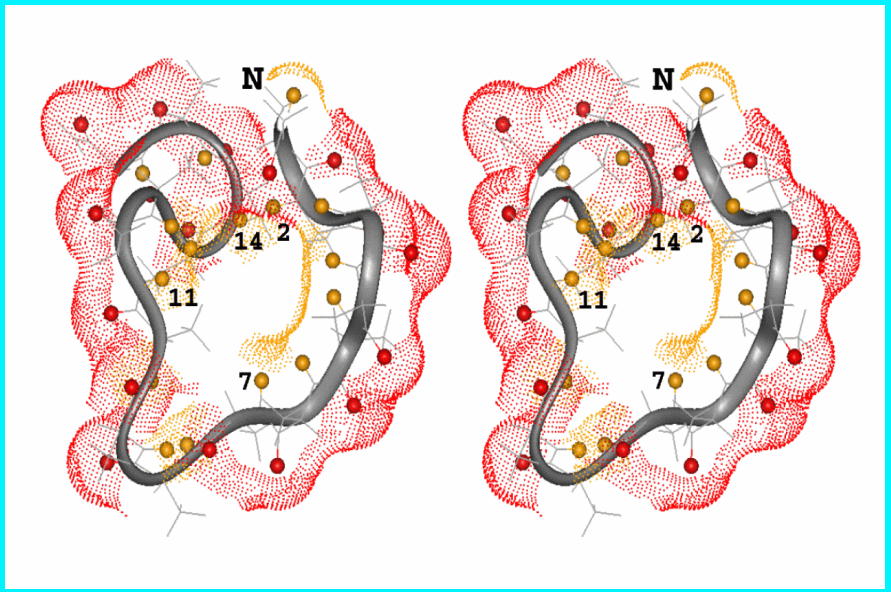
$F_{vib}(\mathbf{r}_p)$ is the conformational entropy contribution;

$F_{cav}(\mathbf{r}_p)$ is the free energy associated with cavity creation;
$F_{solv}(\mathbf{r}_p)$ is the free energy associated with the polarization of the aqueous solution;
$F_{ionz}(\mathbf{r}_p, pH)$ is the free energy associated with the ionization state of the particular conformation of the polypeptide.

APPLICATIONS

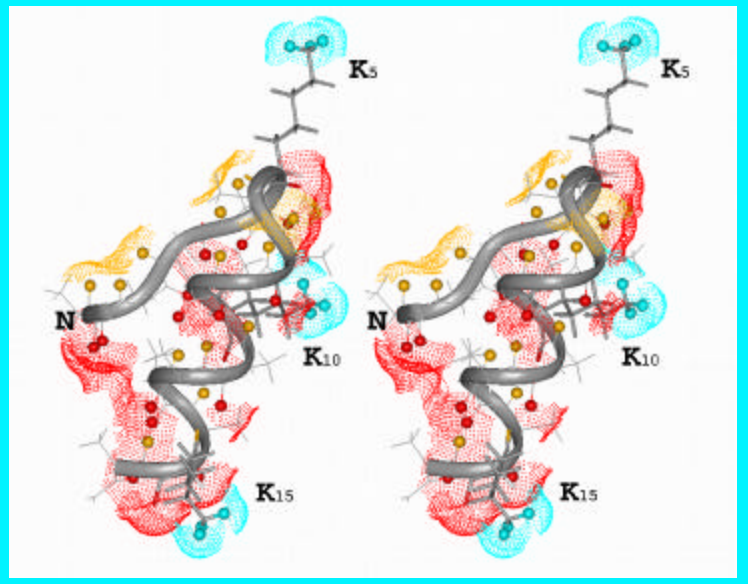
➢The relative contributions of lysine and alanine residues to the stability of α-helices of copolymers of these two residues was investigated through conformational energy calculations on several hexadecapeptides at several pH's.

➢The helix contents were found to depend strongly on the lysine content, in agreement with recent experimental results of Williams *et al*., 1998.

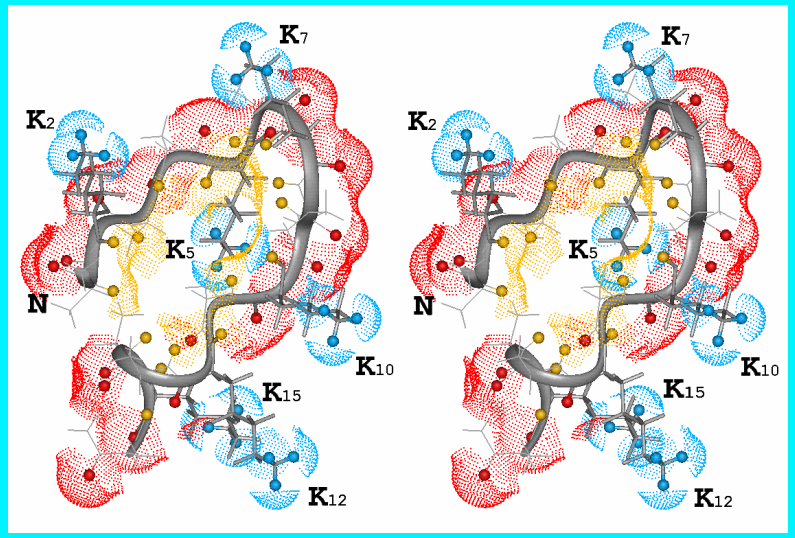


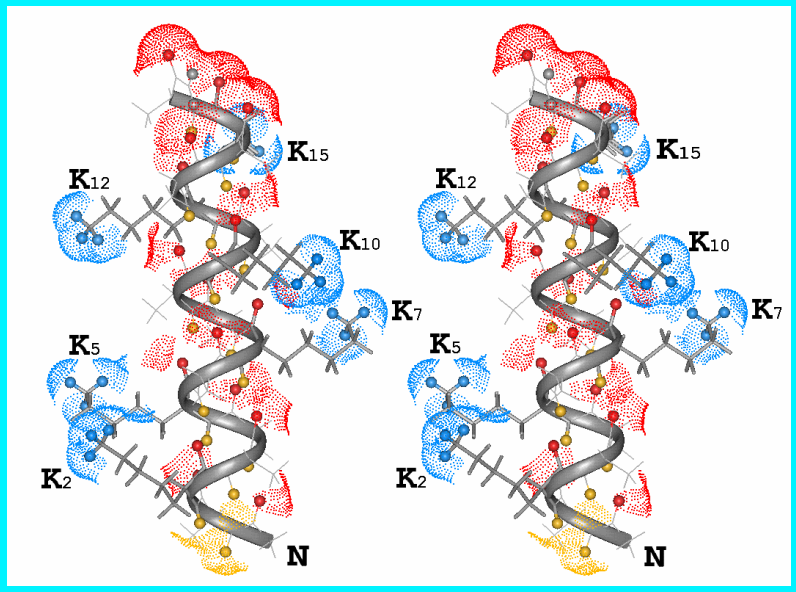
➢In the lowest-energy conformation of a hexadecapeptide containing 3 lysine residues at pH 6, the lysine side chains are preferentially hydrated.

➢This decreases the hydration of the backbone CO and NH groups, forcing them to form hydrogen bonds with each other in the helical conformation.



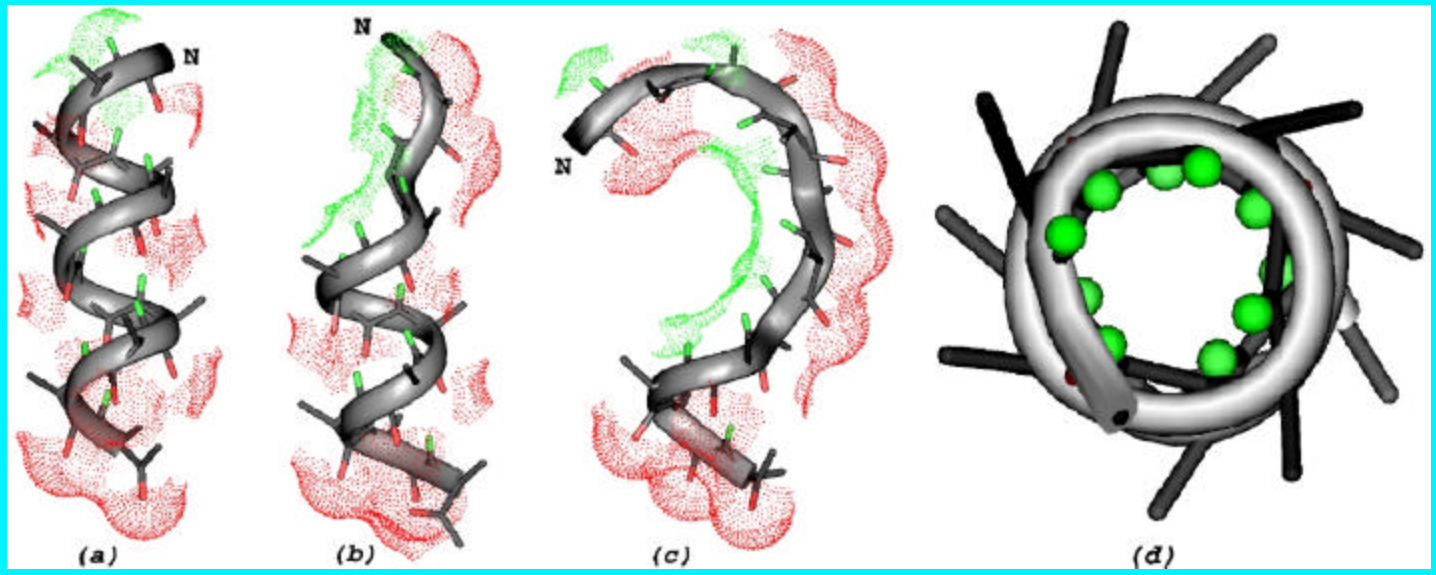
The lowest-energy conformation for a hexadecapeptide containing 6 lysine residues at pH 6 shows a close proximity between the NH_3^+ groups of the lysine side chains, a feature that was previously observed in calculations of short alanine-based oligopeptides.

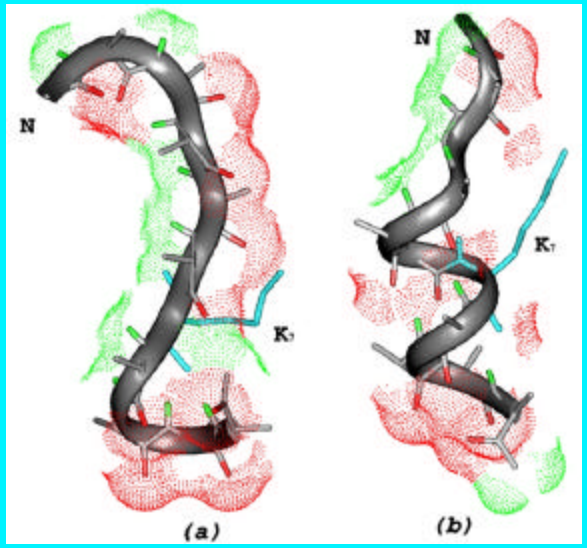




Lowest-energy conformations of ALA_{10} at dielectric constants e=2, e =40, and e =80

Only a fraction of the backbone HN groups, not involved in intramolecular hydrogen bonds, remain exposed to the solvent when ε = 40.





➢Analysis of the sequences with a single lysine or glutamine residue: effect of altering the charge distribution on the side chain and its influence on the conformational preference of the peptide.

➢A single lysine or glutamine residue influences the solvation preference of the backbone CO and NH groups in such a manner as to shift the conformational preference toward the helical conformation in agreement with experimental evidence Kallenbach *et al.*,2000.

The simulations show that:

- Alanine is not a strong helix-forming residue.

- The helix-coil equilibrium of an all-alanine peptide can be shifted toward more compact structures such as the α-helical conformation in two different ways:

  (a) by introducing charged or highly soluble polar residues into the sequence, and

  (b) by lowering the dielectric constant of the solvent.

- Both processes lead to the same effect: the solvation of the CO and NH groups of alanine-based polypeptide is affected by forcing the polar groups to interact more weakly with the solvent and strongly among themselves, leading to the formation of a net of internal hydrogen bonds.

- The α-helical fragments are very characteristic of these compact conformations.

CTC

Collaborators

Cezary Czaplewski

Jooyoung Lee

Adam Liwo

Jaroslaw Pillardy

Harold A. Scheraga

Jorge A. Vila

- Faculty of Chemistry, University of Gdansk,

- Baker Laboratory, Cornell University

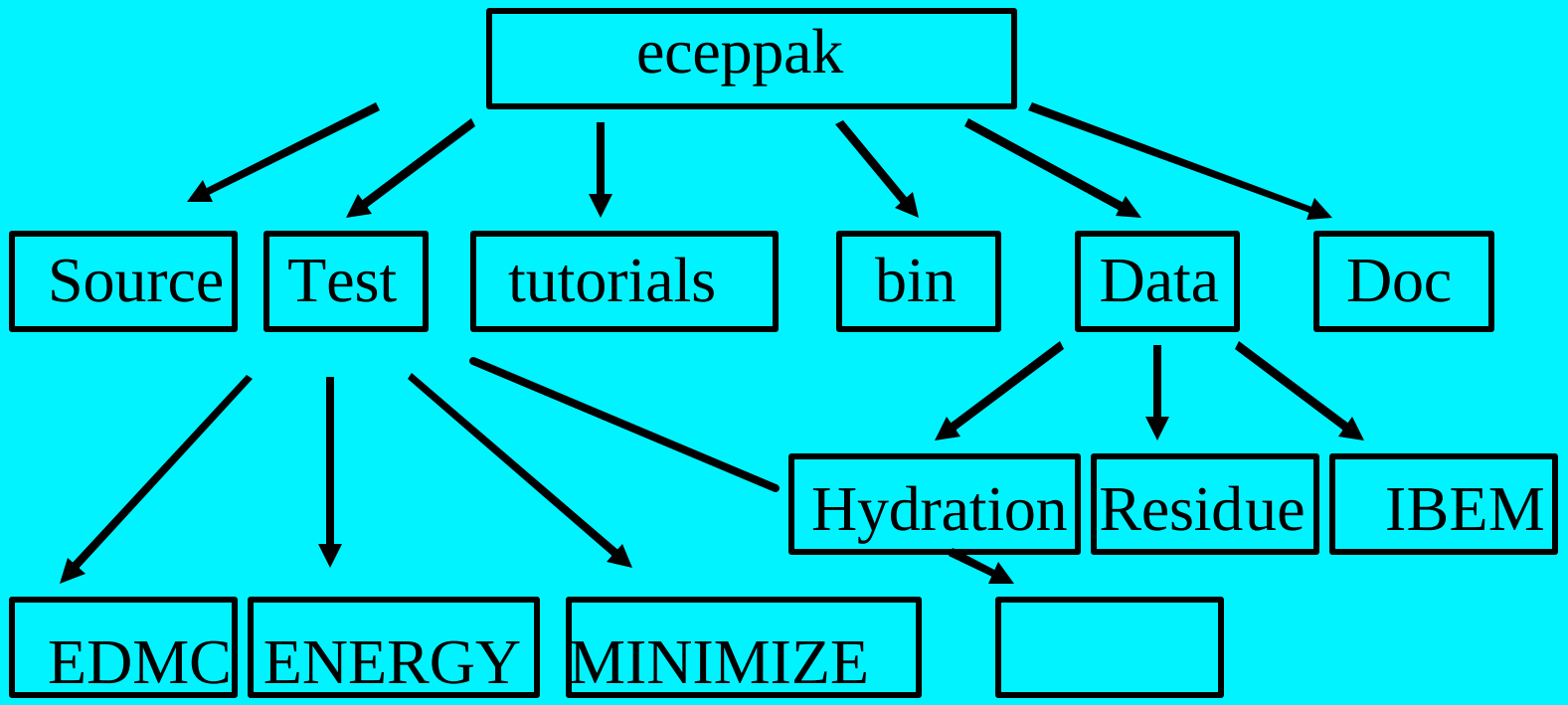
-Cornell Theory Center, Cornell University

-Universidad Nacional de San Luis

CTC

The ECEPPAK Program

Directory structure



Compiling and Running the Program Under WINDOWS

VERY IMPORTANT:

When using the TC computers, some initialization is necessary to access the correct compiler.
Generate a DOS WINDOWS with “command prompt”

To access CYGWIN, type: ***call "H:\CTC Tools\setup_cygwin.bat”***

To obtain access to the compaq compiler, type : ***call setup_compaqf.bat***

Compiling Program Under WINDOWS

The source code of the program, written in FORTRAN 77, is located in the directory C:\eceppak\Source

Makefiles: Makefile-NT.mak and Makefile-NT-MPI to generated the executable for serial and parallel versions, respectively, are provided.

Usage: **make –f *makefile***

Where *makefile* is the appropriate “make” file

CTC

Running the program under WINDOWS

The Windows version of the program can be run interactively or in batch mode in a WINDOWS cluster.

The script ***i.bat*** is used to run the program interactively.

i.bat assigns names of input files, set environment variables used by the ECEPPAK program and launches the executable.

Running the program under WINDOWS

The four scripts used in batch mode:
r.bat, **cp1.bat**, **nwr.bat** and **rmtmp.bat**

r.bat sends the job to the batch queue. Should be edited to define the queue, type and number of nodes and processors to be used, and the input files.

cp1.bat used for copying some input files to t: in each node. This avoids opening large number of files through the network. The name of the directory in t: must be provided, although the default should work fine.

nwr.bat runs the program. No need to change it except that it should point to the same t: directory created by cp1.bat.

rmtmp.bat used for removing the directory and files created in t:

Running the program: *"i.bat"*

```
set ECEPPAK_RES=C:\eceppak\Data\Residue
set ECEPPAK_HYD=C:\eceppak\Data\Residue
set ECEPPAK_BIN=C:\eceppak\bin
set ECEPPAK_IBEM=C:\eceppak\Data\IBEM

set RUNTYP=%1
set INPF=%2
set OUTF=%3
set ANAF=%4
set BNDF=%5

set SOLVDBS=%ECEPPAK_HYD%
set RSDBS=%ECEPPAK_RES%
set SOLVFILE=%ECEPPAK_HYD%\srfopt.set
set FXVOLFILE=%ECEPPAK_HYD%\fix_vol_tst.parm
```

Running the program: *"i.bat" (cont)*

```
set RSDATA=%ECEPPAK_RES%\rsdata
set HLABELFILE=%ECEPPAK_RES%\nmr_protons.data
set DIHDEFILE=%ECEPPAK_RES%\dihedral_def

set IBEMCHG=%ECEPPAK_IBEM%\resibem

set TGTARS=TARGET.tgtars
set RESTFILE=edmc.rstart
set RANDFILE=edmc.rndom

%ECEPPAK_BIN%\eceppak-NT.exe
```

CTC

ECEPPAK: input files:

****.inp*** instruction file for ECEPPAK

outo.* set of dihedral angles for evaluation of multiple conformations

directory eceppak/Data/Residue

rsdata: Residue data file

nmr_protons.data: names with identification of protons (for NMR purposes)

dihedral_def: description of atoms that defines the variable dihedral angles

directory eceppak/Data/Hydration

solvation parameters: ***srfopt.set, volume.set***, etc.

CTC

ECEPPAK: The *.inp file

- Used for **passing instructions** to the program.
- The **parser** reads and interprets the first **78** characters of a line.
- Symbols **#** and **!** indicate a comment (rest of line is ignored).
- **Instructions** for a given procedure are included into **Data Groups**.
- A data group is identified by a **main keyword** which contains the symbol '**$**' as the first character, e.g., $CNTRL, $SEQ, $EDMC.
- Three Data Groups must be included in all input files: **$CNTRL**, **$SEQ** and **$GEOM**.

CTC

ECEPPAK: tutorial.one

Generation of a polyalanine chain and evaluation of its conformational energy

Use the **$CNTRL** Data Group for define the type of run:

```
$CNTRL
runtyp = energy
$end
```

- Use the **$SEQ** data group to define the sequence, and end groups

```
$SEQ
4
1 1 1 1 1 1 1 1 1 1
15
$END
```

ECEPPAK: tutorial.one (cont).

- Include the **$GEOM** data group, with the set of dihedral angles defining the conformation.

```
$GEOM
 180.000 180.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 180.000
$END
```

- Save the file using "**inp**" as the extension, e.g., ten_ala.inp.

ECEPPAK: tutorial.one (cont).

Run the eceppak job by typing:

i.bat ENERGY ten_ala TEN_ALA r r 1>>log 2>err

Output Files:

main_out.TEN_ALA contains the results of the energy evaluation.

Other files: log, err

ECEPPAK: tutorial.one (cont).

To run eceppak in BATCH Mode, edit ***r.bat*** and ***cp1.bat*** files and set the proper parameters;

For submitting the job by typing:

qsub start

Output Files:

main_out.000TEN_ALA contains the results of the energy evaluation.

Other files: log, err

ECEPPAK: tutorial.two

Minimizing the energy and writing coordinates

- Use the **$CNTRL** Data Group to define the type of ECCEPAK run and request the coordinates of the final conformation

```
$CNTRL
runtyp = minimize        ! run type is a minimization
PRINT_CART               !  print Cartesian coordinates
OUTFORMAT   =PDB  !  format of the Cartesian file is PDB
FILE  = ala10M           !    prefix of the PDB file is  ala10M
res_code= three_letter
$END
```

- Use the **$SEQ** data group to define the sequence. Note that the same sequence used in tutorial.one is now specified using a three-letter code.

```
$SEQ
ACE
ALA ALA ALA ALA ALA ALA ALA ALA ALA ALA
NME
$END
```

ECEPPAK: tutorial.two (cont).

-Finally, let's include the same set of dihedral angles as in tutorial.one, using the **$GEOM** data group.

```
$GEOM
 180.000 180.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 -66.000 -40.000 180.000  60.000
 180.000
$END
```

- Save the file using the "inp" extension, e.g., ten_ala_min.inp.

ECEPPAK: tutorial.two (cont).

To run eceppak, type:

i.bat MINIMIZE ten_ala_min TEN_ALA_MIN r r 1>>log 2>err

Output Files:

- main_out.TEN_ALA_MIN with the results of the minimization procedure.
- outo.TEN_ALA_MIN a file containing the final (minimized) conformation with the sequence (in ECEPP format) followed by the list of dihedral angles.
- A10M.pdb files containing the Cartesian coordinates of the the minimized conformations.

CTC

ECEPPAK: tutorial.three

Conformational search using the EDMC method

- Use the **$CNTRL** Data Group to define the type of run (EDMC) and request the coordinates of the accepted conformations:

```
$CNTRL
runtyp = edmc              ! run type is a edmc run
PRINT_CART                 !  print Cartesian coordinates
OUTFORMAT   =PDB   !  format of the Cartesian file is PDB
FILE  = A10edmc          !   prefix of the PDB file is  A10edmc
res_code= one_letter
$END
```

- Use the **$SEQ** data group to define the sequence. Note that the same sequence as in tutorial.one is specified using a one-letter code.

```
$SEQ
A
AAAAAAAAAA
C
$END
```

ECEPPAK: tutorial.three (cont).

•Let's include a set of dihedral angles specifying an extended conformation as input, using the **$GEOM** data group.

```
$GEOM
 180.000 180.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
-160.000-140.000 180.000  60.000
 180.000
$END
```

ECEPPAK: tutorial.three (cont).

The conformational search protocol is defined through specific keywords included in the data group **$EDMC**.

```
$EDMC
MAXIT=20              ! Length of the run
SEED= -5555           ! seed for the random number generator
TEMP= 300            !  Temperature
THERMAL_SHOCK T_UP = 5000 ! sudden jump in the temperature
RAND_TO_ELEC=0.3 ! ratio of randomly- to electrostatically-generated
                          ! conformations
$END
```

-Save the file using the "inp" extension, e.g., ten_ala_edmc.inp.

CTC

ECEPPAK: tutorial.three (cont).

You run eceppak by typing:

i.bat EDMC ten_ala_edmc TEN_ALA_EDMC x x 1>>log 2>err

Output Files:

- main_out.TEN_ALA_EDMC; results of the conformational search procedure.
- outo.TEN_ALA_EDMC; all the conformations accepted by the Monte Carlo procedure. Each conformation is specified by its sequence (in ECEPP format) followed by the list of dihedral angles.
- clustr.IONZ containing the dihedral angles of the lowest-energy conformation representative of each cluster.
- A10edmc###.pdb files (### represents the number of accepted conformation) containing the Cartesian coordinates of the accepted conformations.

CTC

ECEPPAK: tutorial.three (cont).

To visualize the trajectory of local minima followed by the EDMC method, use the molmol.csh script in your local machine

Run

tcsh molmol.csh

ECEPPAK: tutorial.four

Considering the ionization equilibrium in a conformational search

We use here a four-residue sequence to show how to carry out this type conformational search. The sequence is short to speed up the computations.

1- Generate an input file with suffix "inp" (i.e., ionz.inp) that contains the instructions for ECEPPAK.

2.- Include in ionz.inp the $CNTRL Data Group to define the type of run (EDMC).

```
$CNTRL
runtyp = edmc
$END
```

ECEPPAK: tutorial.four (cont)

3.- Include the $SEQ data group with the sequence. As a test, let's consider the four residue-sequence LYS-HIS-LYS-GLU with the blocking end groups at the N- and C-terminus. Let's use ECEPP numbers to specify it.

The $SEQ data group should read,

```
$SEQ
4
09 07 09 04
15
$END
```

ECEPPAK: tutorial.four (cont)

4.-For this run, we are going to use randomly-generated dihedral angles for the initial conformation. Consequently, the $GEOM data group can be filled with arbitrary values. Even empty lines (one per residue or end-group) works.

$GEOM

$END

ECEPPAK: tutorial.four (cont)

5.- The conformational search protocol is defined through a set of specific keywords included in the data group $EDMC. Most of the EDMC keywords (see manual) are assigned default values. We enter a few of them to produce a short run:

```
$EDMC
MAXIT=5
SEED= -677676
TEMP= 300
THERMAL_SHOCK T_LOW=200 T_UP=8000
RAND_START
OMEGA_180
$END
```

CTC

ECEPPAK: tutorial.four (cont)

7.- To carry out the the calculation of the ionization equilibrium, we need to add to additional data groups to our ionz.inp file: $FFIELD and $IBEM_SIMS. The $FFIELD is used to specify the pH at which the ionization equilibrium should be calculated, e.g. pH 7,

```
$FFIELD
  PH=7.0
$END
```

ECEPPAK: tutorial.four (cont)

8.- The $IBEM_SIMS data group is used to input the parameters used by the IBEM algorithm. The IBEM algorithm is used to solved the Poisson-Boltzmann equation.
The following is a list of the parameters and some standard values:

```
$IBEM_SIMS
PROBE_RADIUS=1.4   ! radius of solvent probe sphere;
                   !default = 1.4
DETAIL_OUTPUT=  0      #  values are= 0, 1, 2, 3
                  ! 0 (no output), 1 (some), 2 (more), 3 (max), default = 0
DOT_DENSITY_L=1.25    ! default = 1.0

EPS_SOL= 80.0    ! Dielectric constant of solvent; default = 80
EPS_MOL = 2.0       ! Dielectric constant of molecule interior; default = 2.0
RRLOCK =  5.0       ! radius of local surface element; default = 5.0
```

ECEPPAK: tutorial.four (cont)

```
RRINTERM =8.0              ! radius of intermediate surface element; default = 8.0
NITER_MAX = 50             ! maximum number of iterations; default = 50
DELTA_SIGMA = 0.0001 ! convergence criterion;  default = 0.0001

LINMETH =  BCG       !  'BCG' or 'G-S' default = 'BCG'
                     ! 'BCG' linear system is solved using a biconjugate
                     !   gradient method
                     ! 'G-S' linear system is solved by GAUSS-SEIDEL method
RRLOCK_EN =  10.0  ! radius of calculation by small surface element;
                     !  default = 10.0
ICHARATOM = 0        ! 1  0 sigma is calculated for each charge
ICHARGROUP= 1        ! 0  1 charges as a collection of charged groups
IONECENTR=1          ! 0  1  all charges are considered as one group
$END
```

ECEPPAK: tutorial.four (cont)

6.- Save the file and run ECEPPAK.

To run eceppak edit ***startpH*** and ***receppakpH.csh*** files and set the proper parameters as described previously.

Submit the job by typing:
qsub ***starpH***

ECEPPAK: tutorial.four (cont)

7.- As output, the program writes three different type of files:

• main_out.000IONZ with a description of the results of theconformational search procedure.

• outo.000IONZ containing the dihedral angles of all the conformations accepted by the Monte Carlo procedure.

•(c) clustr.IONZ containing the dihedral angles of the lowest-energy conformation representative of each cluster.

•The file main_out.000IONZ provides the energy components of all the accepted conformations and a description of the state of ionization of the different ionizable residues.